AD-A033 598 JOHNSON MATTHEY AND MALLORY LTD TORONTO (ONTARIO) F/G 12/1
ESTIMATING VARIANCE IN THE PRESENCE OF OUTLIERS.(U)
1976 D E JOHNSON DAHC04-75-G-0168
UNCLASSIFIED ARO-13372.1-M NL

1 OF 1
AD
A033598

END
DATE
FILMED
2-77

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER ARO 13372.1-M	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) Estimating σ^2 in the Presence of Outliers. Variance	5. TYPE OF REPORT & PERIOD COVERED Final Report, 1 Jun 75 - 31 Aug 76
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dallas E. Johnson	8. CONTRACT OR GRANT NUMBER(s) DAHC04-75-G-0168 NEW
9. PERFORMING ORGANIZATION NAME AND ADDRESS Kansas State University	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709	12. REPORT DATE 1976
	13. NUMBER OF PAGES 27
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) 12 29p.	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DDC RECEIVED DEC 16 1976 F

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Outliers	Distributions
Estimators	Random distributions
Tests for outliers	Random samples
Multiple outliers	Unbiased estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

New estimators for variance σ^2 have been proposed and some of their properties have been examined. The estimators V_k ~~and V_k^*~~ are quite simple to calculate and give a good range in which the variance σ^2 will probably fall. However, tables for the divisors V_k have not yet been generated for any cases other than when the sample size is 10.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified 192 250 LB

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ADA033598

(12)

FINAL REPORT

1. ARO PROPOSAL NUMBER: 13372M

2. PERIOD COVERED BY REPORT: June 1, 1975 - August 31, 1976

3. TITLE OF PROPOSAL: Estimating σ^2 in the Presence of Outliers

4. CONTRACT OR GRANT NUMBER: DAHC 04-75-G-0168

5. NAME OF INSTITUTION: Kansas State University

6. AUTHOR OF REPORT: Dallas E. Johnson

7. LIST OF MANUSCRIPTS SUBMITTED OR PUBLISHED UNDER ARO SPONSORSHIP DURING THIS PERIOD, INCLUDING JOURNAL REFERENCES:

   Estimating σ^2 in the Presence of Outliers - to be submitted.

8. SCIENTIFIC PERSONNEL SUPPORTED BY THIS PROJECT AND DEGREES AWARDED DURING THIS REPORTING PERIOD:

   Robert Hughes - no degree awarded

   Steve McGuire - no degree awarded

Estimating σ^2 in the Presence of Outliers

submitted by

Dallas E. Johnson

1. INTRODUCTION

Let $x_1, x_2, \ldots, x_n$ be independent, normally distributed observations with common variance σ^2 and (under H_o) with common mean μ. Many authors, [1,3,4,5,6,9,10,11, 12,13] have been interested in the problem of testing for outlying observations, (i.e., discovering which of the observations, if any, come from a population with a mean different than μ). Grubs [6] gives an excellent review of the literature up to 1967. He discusses procedures which are recommended for the following situations:

1) σ^2 is known

2) σ^2 is not known but an estimate of σ^2 is known which is independent of the sample observations.

3) The only available information is the sample observations.

For (3) above, Grubbs discussed procedures for the cases when there are:

(i) One spurious observation on the high (low) side.

(ii) Two spurious observations on the high (low) side.

(iii) One spurious observation on the high side and one on the low side.

Grubbs remarks that "these techniques are not generally recommended for repeated rejection, since if several outliers are present in the sample the detection of one or two spurious values may be 'masked' by the presence of other anomalous observations". In other words, if there are more than two spurious observations, the procedures given will not be apt to detect the one or two "most" spurious observations.

David and Paulson [2] discuss the performance of several of the tests for outliers under the alternative that there is one spurious observation. McMillan and David [10] discuss the performance of the two procedures when σ^2 is known and under

the alternative that there are two spurious observations. McMillan [9] compares the performance of three procedures when σ^2 is not known and under the alternative that there are two spurious observations.

Tietjin and Moore [12] discuss procedures for detecting outliers when k of the n observations may be considered suspect with $k<n/2$. They state that "suspected observations sometimes form subgroups; i.e., several values are closer to each other than they are close to the bulk of the observations. This phenomenon makes sequential procedures insensitive. It has been called the masking effect." They propose two statistics designed to detect multiple outliers for the case when an independent estimate of σ^2 does not exist.

Guttman and Smith [7] have been the only authors who have considered the problem of estimating σ^2 when spurious observations are present. They considered the problem when there is at most one spurious observation and when μ is unknown, only for a sample of size $n=3$.

The estimation of σ^2 is a very inportant problem. If the data analyst has a "good" estimate of σ^2, he could use it to help judge which, if any, of the observations are outliers and how many outliers there are. Almost all of the tests for outliers are based on σ^2 or some estimate of σ^2, thus it is clear just how important it is to have a "good" estimate of σ^2. New estimators of σ^2 are being proposed here and some of their properties are being examined.

2. NOTATION AND MOTIVATION

Let $x_1, x_2, \ldots, x_n$ be independent normally distributed observations with a common variance σ^2. Suppose that k_1 of the observations come from a population with mean $\mu+\lambda$ and $k_2 = n-k_1$ of the observations come from a population with mean μ. If $k_1 = 0$, the uniformly minimum variance unbiased estimator of σ^2 is given by $s^2 = \Sigma(x_i-\bar{x})^2/(n-1)$. Note that one can also calculate s^2 by the formula

$s^2 = \Sigma_{i=1}^n \Sigma_{j=1}^n (x_i-x_j)^2/2n(n-1)$. This is an interesting way to look at s^2 for if one of the observations were an outlier, say x_q, then the absolute

differences $|x_i-x_q|$, $i\neq q$ would be large in comparison to the absolute differences not involving x_q. If one knew that x_q was a spurious observation and the only spurious observation, the uniformly minimum variance unbiased estimator of σ^2 would be given by

$$\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} c_{ij}(x_i-x_j)^2/2(n-1)(n-2)$$

where $c_{ij}=0$ if $i=q$ or if $j=q$ and $c_{ij}=1$ otherwise.

More generally, suppose $x_{q_1}, x_{q_2}, \ldots, x_{q_{k_1}}$ come from the population with mean $\mu+\lambda$ and $x_{p_1}, x_{p_2}, \ldots, x_{p_{k_1}}$ come from the population with mean μ. Let $A=\{q_1,q_2,\ldots,q_{k_1}\}$ and $B=\{p_1,p_2,\ldots,p_{k_2}\}$. Then the uniformly minimum variance unbiased estimator of σ^2 is

$$\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}c_{ij}(x_i-x_j)^2/2[k_1(k_1-1)+k_2(k_2-1)]$$

where $c_{ij} = 1$ if $i \in A$, $j \in A$ or if $i \in B$, $j \in B$ and $c_{ij} = 0$ if $i \in A$, $j \in B$ or if $i \in B$, $j \in A$.

An intuitive feeling for what could be done is to exclude from the sum, $\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}(x_i-x_j)^2$, those terms for which $\left|x_i-x_j\right|$ is quite large, say larger than some prespecified constant c. That is, it seems reasonable that a "good" estimator of σ^2 can be based on the statistic

$$Q_c = \Sigma_{i=1}^{n}\Sigma_{j=1}^{n}c_{ij}(x_i-x_j)^2 \text{ where } c_{ij} = 1 \text{ if } |x_i-x_j|\le c \text{ and } c_{ij} = 0 \text{ if } |x_i-x_j|>c.$$

This is similar to a successful procedure introduced by Johnson and Graybill [8] for estimating σ^2 and locating outliers in the two-way classification model.

Of many methods of choosing c, there are two which are appealing. First, if the experimentor has some prior knowledge about the relative size of σ^2, he could choose c accordingly. That is, c could be chosen as some multiple of σ such as $c = 2\sqrt{2}\,\sigma$. This particular value would eliminate those differences from the sum which are more than two standard deviations apart.

A second method of choosing c is to choose c such that a given percentage of the absolute differences are excluded from the sum. To do this let $u_{ij} = |x_i-x_j|$ for $i<j=1,2\ldots, n$ and let $u_{(1)}\ge u_{(2)}\ge\ldots\ge u_{(n(n-1)/2)}$ be the ordered absolute differences.

One can then let $c = u_{(n)}$ for some r, $1 \le r \le n(n-1)/2$. If there are k_1 outliers in the data, there will be $k_1 k_2$ comparisons which should be excluded from the sum. Thus, the obvious choices for r which have been considered are $1\cdot(n-1)$, $2\cdot(n-2)$, $3\cdot(n-3)$,... .

In the next section new estimators of σ^2 are proposed and some of their properties are discussed. In some cases their properties are examined by exact methods and in other cases the properties are examined by Monte-Carlo methods.

3. ESTIMATORS OF σ^2

3.1 Method 1 Estimators

Let $Q_c = \Sigma_{i=1}^{n}\Sigma_{j=1}^{n} c_{ij}(x_i - x_j)^2/2n(n-1)$ where $c_{ij} = 1$ if $|x_i - x_j| \le c$ and $c_{ij} = 0$ if $|x_i - x_j| > c$ and where c is some prespecified constant.

First the expected value of Q_c is obtained. The expected value of Q_c is obtained by making repeated applications of the following lemma.

Lemma 3.1 Suppose $W \sim N(\delta, \Theta^2)$. Let G be the random variable defined by $G=1$ if $|W| \le c$ and $G=0$ if $|W| > c$. Then

$$E(GW^2) = \Theta^2[(1+\delta^2/\Theta^2)(1-z((c-\delta)/\Theta)-z((c+\delta)/\Theta) - \frac{1}{\sqrt{2}}(\frac{c+\delta}{\Theta}\exp\{-(c-\delta)^2/2\Theta^2\} + \frac{c-\delta}{\Theta}\exp\{-(c+\delta)^2/2\Theta^2\}]$$

where $z(a) = \int_a^\infty (1/\sqrt{2\pi})\exp\{-x^2/2\}dx$.

Proof:

$$E(GW^2) = \int_{-c}^{c} w^2 f_W(w)\, dw$$

$$= \int_{-c}^{c} (w^2/\sqrt{\Theta\, 2\pi})\, e^{-(w-\delta)^2/2\Theta^2}\, dw$$

To evaluate the integral on the right above, let $v = (w-\delta)/\Theta$, then

$$E(GW^2) = \int_{-(c+\delta)/\Theta}^{(c-\delta)/\Theta} \frac{(\Theta v+\delta)^2}{\sqrt{2\pi}} e^{-v^2/2} dv.$$

The result above is obtained from

this by straight forward integration.

The expected value of Q_c is given by the following theorem.

Theorem 3.1 Suppose $x_1, x_2, \ldots, x_n$ are independent normally distributed random variables with a common variance σ^2. Suppose that k_1 of the x_i come from a population with mean $\mu + \lambda$ and the remaining $k_2 = n-k_1$ of the x_i come from a population with mean μ. Then

$$E(Q_c) = \frac{\sigma^2}{n(n-1)} \{(k_1^2 + k_2^2 - n)(1-2z(c/\sigma\sqrt{2}) - (c/\sigma\sqrt{\pi})\, e^{-c^2/4\sigma^2})$$

$$+ 2k_1k_2[(1+\lambda^2/2\sigma^2)(1-z((c+\lambda)/\sigma\sqrt{2}) - z((c-\lambda)/\sigma\sqrt{2}))$$

$$- \frac{1}{2\sigma\sqrt{\pi}} ((c+\lambda)\, e^{-(c-\lambda)^2/4\sigma^2} -(c-\lambda)\, e^{-(c+\lambda)^2/4\sigma^2})]\} \qquad (3.1)$$

Proof: Without any loss of generality it can be assumed that $x_1, x_2, \ldots, x_{k_1}$ have mean $\mu + \lambda$ and $x_{k_1+1}, x_{k_1+2}, \ldots, x_n$ have mean μ. Let $S_1 = \{1,2,\ldots,k_1\}$ and $s_2 = \{k_1+1, k_1+2 \ldots, k_1 + k_2\}$.

Now

$$E(2n(n-1)Q_c) = E[\Sigma_{i=1}^n \Sigma_{j=1}^n c_{ij}(x_i-x_j)^2]$$

$$= \Sigma_{i=1}^n \Sigma_{j=1}^n E[c_{ij}(x_i-x_j)^2]$$

$$= \Sigma_{i,j\varepsilon s_1} E[c_{ij}(x_i-x_j)^2] + \Sigma_{i\varepsilon s_1, j\varepsilon s_2} E[c_{ij}(x_i-x_j)^2]$$

$$+ \Sigma_{i\varepsilon s_2, j\varepsilon s_1} E[c_{ij}(x_i-x_j)^2] + \Sigma_{i,j\varepsilon s_2} E[c_{ij}(x_i-x_j)^2]$$

$$= H_1 + H_2 + H_3 + H_4 \text{ (say).}$$

Each of the terms in H_1 and H_4 where $i \neq j$ have $x_i - x_j \sim N(0, 2\sigma^2)$, thus, according to Lemma 3.1,

$$E[c_{ij}(x_i-x_j)^2] = 2\sigma^2[1-2z(c/\sigma\sqrt{2}) - (c/\sigma\sqrt{\pi})e^{-c^2/4\sigma^2}] = G_1 \text{ (say)}$$

for every term in H_1 and H_4.

Each term in H_2 has $x_i-x_j \sim N(\lambda, 2\sigma^2)$ and each term in H_3 has $x_i-x_j \sim N(-\lambda, 2\sigma^2)$ hence, according to Lemma 3.1,

$$E[c_{ij}(x_i-x_j)^2] = 2\sigma^2[(1+\lambda^2/2\sigma^2)(1-z((c-\lambda)/\sigma\sqrt{2}) - z((c+\lambda)/\sigma\sqrt{2})$$
$$- \frac{1}{\sqrt{2\pi}}\left(\frac{c+\lambda}{\sigma\sqrt{2}} e^{-(c-\lambda)^2/4\sigma^2} + \frac{c-\lambda}{\sigma\sqrt{2}} e^{-(c+\lambda)^2/4\sigma^2}\right)] = G_2 \text{ (say)}$$

for every term in H_3 and H_4. Therefore

$$E[2n(n-1)Q_c] = k_1(k_1-1)G_1 + k_1k_2G_2 + k_1k_2G_2 + k_2(k_2-1)\, G_1$$
$$= (k_1^2 + k_2^2-n)\, G_1 + 2k_1k_2G_2$$
$$= 2\sigma^2\ \{(k_1^2 + k_2^2-n)(1-2z(c/\sigma\sqrt{2}) - (c/\sigma\sqrt{\pi})\, e^{-c^2/4\sigma^2})$$
$$+ 2k_1k_2[(1+\lambda^2/2\sigma^2)(1-z((c-\lambda)/\sigma\sqrt{2})-z((c+\lambda)/\sigma\sqrt{2}))$$
$$- \frac{1}{\sqrt{2\pi}}\left(\frac{c+\lambda}{\sigma\sqrt{2}} e^{-(c-\lambda)^2/4\sigma^2} + \frac{c-\lambda}{\sigma\sqrt{2}} e^{-(c+\lambda)^2/4\sigma^2}\right)]\}$$

The desired result follows from this.

The following two corollaries are useful in choosing an estimator of σ^2 based on Q_c.

<u>Corollary 3.1</u> If $\lambda=0$ or equivalently if $k_1=0$, then

$$E[Q_c] = \sigma^2(1-2z(c/\sigma\sqrt{2}) - (c/\sigma\sqrt{\pi})\, e^{-c^2/4\sigma^2})$$

<u>Corollary 3.2</u>

$$\lim_{\lambda\to\infty} E[Q_c] = \frac{k_1^2 + k_2^2 - n}{n(n-1)} [1-2z(c/\sigma\sqrt{2}) - (c/\sigma\sqrt{\pi})\, e^{-c^2/4\sigma^2}]\, \sigma^2.$$

<u>Theorem 3.2</u> Let $S^2 = \sum_{i=1}^n (X_i-\bar{X})^2/(n-1)$. Under the assumption given in Theorem 3.1,

$$E(S^2) = \sigma^2 + \frac{k_1 k_2}{n(n-1)} \lambda^2.$$

Tables 1 and 2 give values of $E(Q_c/\sigma^2)$ and $E(S^2/\sigma^2)$ for some different values of n, k_1, λ, and c.

1. $E(Q_c/\sigma^2)$ and $E(S^2/\sigma^2)$ when $n = 10$.

c	$\sqrt{2}\sigma$			$2\sqrt{2}\sigma$			$3\sqrt{2}\sigma$			$4\sqrt{2}\sigma$			$E(S^2/\sigma^2)$		
k_1	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$\lambda=0$	.199	.199	.199	.739	.739	.739	.971	.971	.971	.999	.999	.999	1.0	1.0	1.0
$\lambda=3\sigma$	.170	.146	.132	.761	.779	.792	1.390	1.716	1.949	1.782	2.392	2.827	1.9	2.6	3.1
$\lambda=6\sigma$	.159	.128	.106	.598	.488	.410	.916	.874	.843	1.696	2.239	2.626	4.6	7.4	9.4
$\lambda=9\sigma$	.159	.128	.106	.591	.476	.394	.777	.627	.519	.823	.687	.590	9.1	15.4	19.9
$\lambda=12\sigma$	.159	.128	.106	.591	.476	.394	.777	.626	.518	.799	.644	.533	15.4	26.6	34.6
$\lambda=\infty$	.159	.128	.106	.591	.476	.394	.777	.626	.518	.799	.644	.533	∞	∞	∞

2. $E(Q_c/\sigma^2)$ and $E(S^2/\sigma^2)$ when n = 25.

c	$\sqrt{2}\sigma$			$2\sqrt{2}\sigma$			$3\sqrt{2}\sigma$			$4\sqrt{2}\sigma$			$E(S^2/\sigma^2)$		
k_1	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$\lambda=0$	.199	.199	.199	.739	.739	.739	.971	.971	.971	.999	.999	.999	1.0	1.0	1.0
$\lambda=3\sigma$	.187	.177	.167	.748	.756	.764	1.138	1.292	1.432	1.312	1.600	1.861	1.375	1.690	1.9[illegible]
$\lambda=6\sigma$	.183	.168	.155	.683	.631	.584	.949	.929	.911	1.278	1.534	1.766	2.44	3.76	4.9[illegible]
$\lambda=9\sigma$	.183	.168	.155	.679	.625	.576	.893	.822	.758	.929	.864	.806	4.24	7.21	9.9[illegible]
$\lambda=12\sigma$	.183	.168	.155	.679	.625	.576	.893	.822	.757	.919	.846	.779	6.76	12.04	16.8[illegible]
$\lambda=\infty$	.183	.168	.155	.679	.625	.576	.893	.822	.757	.919	.846	.779	∞	∞	∞

Upon examination of Tables 1 and 2 one can see that Q_c is not affected as much by outliers as is the usual estimator of σ^2, S^2. Unfortunately, Q_c usually under estimates σ^2. Examination of these two tables shows that one must be careful not to pick c too small. The most optimal choice appears to be somewhere around $3\sqrt{2}\sigma$ and $4\sqrt{2}\sigma$. Although it would be possible to divide Q_c by a constant to reduce the bias, this will not be done here because other estimators are being proposed which have more desirable properties.

3.2 Method 2 Estimators

Let $u_{ij} = \left|x_i - x_j\right|$ for i<j = 1, 2,..., n and let $u_{(1)} \geq u_{(2)} \geq \cdots \geq u_{(n(n-1)/2)}$ be the ordered values of the u_{ij}. If there is one outlier in the data (i.e. $k_1=1$), it is reasonable to hope that the n-1 differences between the outlier and the remaining observations will be larger than the (n-1)(n-2)/2 differences between the remaining n-1 observations. Hence, if these n-1 differences are removed from the sum of all squared differences, a reasonable estimator of σ^2 would be given by

$$V_1 = (\sum_{i=1}^{n}\sum_{i=j}^{n} u_{1j}^2 - 2\sum_{i=1}^{n-1} u_{(i)}^2)/2(n-1)(n-2)$$

or equivalently by

$$V_1 = (\sum_{i<j} u_{ij}^2 - \sum_{i=1}^{n-1} u_{(i)}^2)/(n-1)(n-2).$$

More generally, for k_1 outliers an intuitive estimator of σ^2 is given by

$$V_{k_1} = (\sum_{i<j} u_{ij}^2 - \sum_{i=1}^{k_1 k_2} u_{(i)}^2)/(k_1(k_1-1) + k_2(k_2-1)).$$

In order to examine the properties of these estimators and to evaluate how "good" these estimators are, 1000 random samples of size 10 were generated for each of several cases. This was done for $k_1 = 1, 2, 3, 5$ and $\lambda/\sigma = 0, 3, 6, 9, 12$. The results are presented in Tables 3-6.

3. The mean, variance, and mean square error of V_1.

k_1	λ/σ	$E(V_1/\sigma^2)$	$Var(V_1/\sigma^2)$	$MSE(V_1/\sigma^2)$
-	0	.5057	.0594	.304
1	3	.7948	.1209	.163
1	6	.9897	.2091	.209
1	9	1.0005	.2549	.255
1	12	1.0116	.2551	.255
2	3	1.2081	.2060	.249
2	6	3.2801	.6686	5.867
2	9	7.0790	1.5713	38.526
2	12	12.6224	3.1477	138.228
3	3	1.6060	.3687	.736
3	6	5.2718	1.4831	19.732
3	9	11.936	3.5253	123.138
3	12	21.5141	6.7534	427.580
5	3	1.9605	.4909	1.413
5	6	7.0175	2.2324	38.442
5	9	16.0074	6.0760	231.299
5	12	29.0490	10.4993	797.248

4. The mean, variance, and mean square error of V_2.

k_1	λ/σ	$E(V_2/\sigma^2)$	$Var(V_2/\sigma^2)$	$MSE(V_2/\sigma^2)$
-	0	.3058	.0233	.505
1	3	.4595	.0475	.340
1	6	.5244	.0683	.294
1	9	.5208	.0761	.306
1	12	.5270	.0756	.299
2	3	.6618	.0765	.191
2	6	.9505	.1996	.202
2	9	1.0159	.2685	.269
2	12	.9914	.2685	.269
3	3	.9264	.1225	.128
3	6	2.5836	.3504	2.858
3	9	5.6753	.9042	22.763
3	12	10.2552	1.7572	87.415
5	3	1.1991	.1812	.221
5	6	4.2833	.8790	11.659
5	9	10.0579	2.7473	84.793
5	12	18.6274	4.8748	315.601

5. The mean, variance, and mean square error of V_3.

k_1	λ/σ	$E(V_3/\sigma^2)$	$Var(V_3/\sigma^2)$	$MSE(V_3/\sigma^2)$
-	0	.2039	.0110	.6448
1	3	.3022	.0226	.5095
1	6	.3374	.0301	.4691
1	9	.3370	.0347	.4743
1	12	.3401	.0335	.4690
2	3	.4276	.0375	.3651
2	6	.5492	.0826	.2858
2	9	.5711	.0913	.2753
2	12	.5587	.0934	.2881
3	3	.5794	.0574	.2343
3	6	.9528	.1761	.1783
3	9	.9805	.2710	.2714
3	12	.9905	.2793	.2794
5	3	.7780	.0727	.1220
5	6	2.3389	.2317	2.0243
5	9	5.1775	.6986	18.1497
5	12	9.5095	1.3469	73.7588

6. The mean, variance, and mean square error of V_5.

k_1	λ/σ	$E(V_5\sigma/^2)$	$Var(V_5/\sigma^2)$	$MSE(V^5/\sigma^2)$
-	0	.1402	.0057	.7449
1	3	.2064	.0115	.6414
1	6	.2278	.0147	.6109
1	9	.2280	.0170	.6129
1	12	.2303	.0162	.6087
2	3	.2908	.0193	.5223
2	6	.3587	.0395	.4509
2	9	.3693	.0414	.4385
2	12	.3629	.0423	.4482
3	3	.3914	.0309	.4013
3	6	.5652	.0790	.2681
3	9	.5603	.0943	.2876
3	12	.5736	.1021	.2839
5	3	.5133	.0367	.2736
5	6	.9392	.1502	.1539
5	9	.9774	.2184	.2189
5	12	.9693	.2042	.2052

Examination of Tables 3, 4, 5, and 6 reveals that:

(i) V_k under estimates σ^2 whenever $k \geq k_1$ and λ is small.

(ii) V_k is approximately unbiased for σ^2 when λ is larger and $k=k_1$.

(iii) V_k over estimates σ^2 whenever $k < k_1$ and λ is not small.

With regard to (ii) above, it is easily shown that $E(V_k) \to \sigma^2$ as $\lambda \to \infty$ when $k=k_1$. If λ is large and the number of outliers is known then V_{k_1} is the "best" estimator of σ^2 that is available.

Let $V_k^* = V_k/\nu_k$ when $\nu_k = \underset{\lambda=0}{E} (V_k/\sigma^2)$. Thus V_k^* will be an unbiased estimator of σ^2 where $\lambda=0$; i.e. when there are no outliers in the data. For those cases when $\lambda \neq 0$ and $k_1 > 0$, V_k^* will be a conservative estimate of σ^2. That is, V_k^* over estimates σ^2. While it is true that S^2, the usual sample variance, is also a conservative estimator of σ^2 it is generally much more conservative than V_k. For comparison purposes Table 7 gives the expected value of V_k^* and S^2 for several alternative models. The expected values of S^2 are exact while those for V_k^* come from the Monte-Carlo Study.

Examination of Table 7 reveals that V_1^* is less biased for σ^2 than S^2 when $k_1=1$ and V_2^* is less biased for σ^2 than both S^2 and V_1^* when $k_1 \leq 2$. Similarly V_3^* is less biased than S^2, V_1^*, and V_2^* for $k_1 \leq 3$ and V_5^* is less biased than S^2, V_1^*, and V_2^*, and V_3^* for $k_1 \leq 5$. Thus from the above it would seem that V_5^* should be the preferred estimator of σ^2 for all of the alternative cases; however, one should first examine the variance and/or the mean square error of these estimators before making a final decision. Since all of these estimators are biased except when $\lambda=0$, only the mean square error of the estimators will be given. Table 8 gives the mean square

7. The mean of S^2, V_1^*, V_2^*, V_3^*, and V_5^*.

K_1	λ/σ	$E(S^2/\sigma^2)$	$E(V_1^*/\sigma^2)$	$E(V_2^*/\sigma^2)$	$E(V_3^*/\sigma^2)$	$E(V_5^*/\sigma^2)$
-	0	1	1	1	1	1
1	3	1.9	1.572	1.503	1.482	1.472
1	6	4.6	1.957	1.715	1.655	1.625
1	9	9.1	1.978	1.703	1.653	1.627
1	12	15.4	2.000	1.723	1.668	1.643
2	3	2.6	2.389	2.165	2.098	2.075
2	6	7.4	6.486	3.109	2.694	2.559
2	9	15.4	13.998	3.323	2.802	2.638
2	12	26.6	24.959	3.242	2.741	2.589
3	3	3.1	3.176	3.030	2.842	2.792
3	6	9.4	10.424	8.450	4.674	4.032
3	9	19.9	23.603	18.561	4.810	3.998
3	12	34.6	42.541	33.540	4.859	4.092
5	3	3.5	3.877	3.922	3.816	3.662
5	6	11.0	13.876	14.009	11.473	6.700
5	9	23.5	31.652	32.895	25.397	6.973
5	12	41.0	57.440	60.922	46.647	6.915

8. The mean square error of S^2, V_1^*, V_2^*, V_3^* and V_5^*.

K_1	λ/σ	$MSE(S^2/\sigma^2)$	$MSE(V_1^*/\sigma^2)$	$MSE(V_2^*/\sigma^2)$	$MSE(V_3^*/\sigma^2)$	$MSE(V_5^*/\sigma^2)$
-	0	.22	.23	.26	.26	.29
1	3	1.43	.80	.77	.78	.81
1	6	14.78	1.73	1.26	1.15	1.14
1	9	69.43	1.95	1.33	1.26	1.26
1	12	213.98	2.00	1.36	1.25	1.24
2	3	3.49	2.73	2.20	2.11	2.14
2	6	44.03	32.71	6.64	4.86	4.44
2	9	213.98	175.09	8.35	5.44	4.79
2	12	666.96	586.34	7.98	5.28	4.68
3	3	5.57	6.17	5.47	4.77	4.78
3	6	74.52	94.62	59.35	17.73	13.22
3	9	365.83	524.69	318.34	21.04	13.78
3	12	1144.12	1752.06	1078.17	21.61	14.76
5	3	7.58	10.19	10.53	9.68	8.96
5	6	104.67	174.52	178.90	115.26	40.14
5	9	516.47	963.33	1047.49	612.03	46.80
5	12	1618.00	3226.57	3644.23	2116.09	45.38

error of these estimators. Note that when $\lambda=0$, $MSE(V_k^*)=VAR(V_k^*)$.

From Table 8 it is interesting to note that when $\lambda=0$, the variances of the estimators V_1^*, V_2^*, V_3^*, and V_5^* are not much larger than the variance of S^2. That is, when there are no outliers in the data one does not lose much efficiency by choosing one of the estimators V_1^*, V_2^*, V_3^*, or V_5^* rather than S^2. This seems to be a small price to pay for the reduction in bias that is made when outliers are in the data.

To summarize the results in this subsection a procedure which might be used when one desires to estimate the variance of a sample of size 10 which may or may not have outliers in it can be described as the following:

(i) Decide on an upper bound, k, for the number of outliers in the sample. If the experimentor is reluctant to do this, he may choose $k=5$; i.e. half of the data are outliers or the data is a mixture of two populations.

(ii) Calculate both V_k and V_k^*.

(iii) Conclude that σ^2 is quite likely to fall somewhere between V_k and V_k^*.

With this new information about σ^2 it may now be possible to decide whether some of the data are spurious observations or not.

4. ESTIMATORS OF σ^2 AND λ

In this section the problem of estimating both σ^2 and λ is considered. Estimators of σ^2 and λ are defined and their properties discussed. Throughout this section it is assumed that k_1, the number of outliers, in the data, is known.

4.1 Two equation estimators

Let S^2 be the sample variance and Q_c be as defined in section 3.1. Now then $E(S^2) = \sigma^2 + k_1 k_2 \lambda^2/n(n-1)$ and $E(Q_c)$ is given by ( ). A pseudo-method of moments estimator of σ^2 and λ can be obtained in the following way:

Find the values of σ^2 and λ, say $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$, such that

$$S^2 = \hat{\sigma}_c^2 + k_1 k_2 \hat{\lambda}_c^2/n(n-1) \qquad \text{and}$$

$$Q_c = g_c(\hat{\sigma}_c^2, \hat{\lambda}_c) \text{ where } g_c(\sigma^2, \lambda) = E(Q_c).$$

Unfortunately, these two equations do not always have a solution. In this case $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ are found so that the sum of the squared residuals is minimized. That is, $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ obtained so that

$$R = [S^2 - \hat{\sigma}_c^2 - k_1 k_2 \hat{\lambda}_c^2/n(n-1)]^2 + [Q_c - g_c(\hat{\sigma}_c^2, \hat{\lambda}_c)] \text{ is minimized.}$$

The procedure that was used to find the values of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ that minimize R is a combination of the Gauss-Newton method and the method of

Steepest Descent for estimating the parameters of non-linear models.

Tables 9-12 give the mean, variance, and mean square error of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ for various values of c, λ, and k_1 from the Monte Carlo study.

Examination of these tables show that $\hat{\lambda}_c$ does a "surprisingly good" job of estimating λ for all choices of c and for each value of k_1. In terms of estimating σ^2, however, a value of c somewhere around $2\sqrt{2}\sigma$ or $3\sqrt{2}\sigma$ seems to give approximately unbiased estimates of σ^2 for most values of λ and k_1. In other words, in order to use $\hat{\sigma}_c^2$ to estimate σ^2, one should be careful in selecting the value of c so that it is neither too small or too large.

9. The mean, variance, and mean square error of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ when $k_1 = 1$.

c/σ	λ/σ	$E(\hat{\sigma}_c^2/\sigma^2)$	$Var(\hat{\sigma}_c^2/\sigma^2)$	$MSE(\frac{\hat{\sigma}_c^2}{\sigma^2})$	$E(\hat{\lambda}_c/\sigma^2)$	$Var(\frac{\hat{\lambda}_c}{\sigma})$	$MSE(\frac{\hat{\lambda}_c}{\sigma})$
$\sqrt{2}$	3	.411	.058	.405	3.803	1.402	2.047
$\sqrt{2}$	6	.345	.025	.454	6.439	1.122	1.315
$\sqrt{2}$	9	.342	.025	.457	9.305	1.105	1.198
$\sqrt{2}$	12	.348	.025	.449	12.227	1.197	1.249
$2\sqrt{2}$	3	1.045	.281	.283	2.405	2.446	2.800
$2\sqrt{2}$	6	1.252	.826	.890	5.624	1.532	1.674
$2\sqrt{2}$	9	1.233	.535	.589	8.785	1.317	1.363
$2\sqrt{2}$	12	1.220	.448	.497	11.845	1.221	1.245
$3\sqrt{2}$	3	1.536	.620	.907	2.425	2.562	2.893
$3\sqrt{2}$	6	1.043	.449	.450	5.833	1.573	1.601
$3\sqrt{2}$	9	1.108	.646	.657	8.880	1.156	1.170
$3\sqrt{2}$	12	1.164	.777	.803	11.879	1.235	1.250
$4\sqrt{2}$	3	1.872	.854	1.614	3.235	2.321	2.376
$4\sqrt{2}$	6	1.213	.603	.648	5.702	2.541	2.630
$4\sqrt{2}$	9	.967	.285	.286	8.971	1.064	1.065
$4\sqrt{2}$	12	1.033	.328	.329	11.938	1.178	1.182

10. The mean, variance, and mean square error of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ when $k_1 = 2$.

c/σ	λ/σ	$E(\hat{\sigma}_c^2/\sigma^2)$	$Var(\hat{\sigma}_c^2/\sigma^2)$	$MSE(\hat{\sigma}_c^2/\sigma^2)$	$E(\hat{\lambda}_c/\sigma)$	$Var(\hat{\lambda}_c/\sigma)$	$MSE(\hat{\lambda}_c/\sigma)$
$\sqrt{2}$	3	.536	.204	.390	3.239	1.232	1.289
$\sqrt{2}$	6	.569	.028	.459	6.240	.631	.687
$\sqrt{2}$	9	.343	.029	.473	9.184	.598	.632
$\sqrt{2}$	12	.333	.223	.627	12.123	.634	.649
$2\sqrt{2}$	3	1.243	.866	.925	2.509	1.501	1.742
$2\sqrt{2}$	6	1.328	1.636	1.744	5.730	1.028	1.101
$2\sqrt{2}$	9	1.254	.584	.649	8.879	.701	.716
$2\sqrt{2}$	12	1.209	.473	.517	11.927	.655	.660
$3\sqrt{2}$	3	1.516	.899	1.165	2.599	1.286	1.447
$3\sqrt{2}$	6	1.055	.861	.864	5.882	1.010	1.024
$3\sqrt{2}$	9	1.146	.773	.794	8.921	.621	.627
$3\sqrt{2}$	12	1.171	1.043	1.072	11.934	.677	.681
$4\sqrt{2}$	3	2.291	1.490	3.158	2.469	1.605	1.887
$4\sqrt{2}$	6	1.398	1.635	1.793	5.688	1.903	2.000
$4\sqrt{2}$	9	.965	.257	.259	8.983	.590	.590
$4\sqrt{2}$	12	1.010	.325	.326	11.974	.636	.637

11. The mean, variance, and mean square error of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ when $k_1 = 3$.

c/σ	λ/σ	$E(\hat{\sigma}_c^2/\sigma^2)$	$Var(\hat{\sigma}_c^2/\sigma^2)$	$MSE(\hat{\sigma}_c^2/\sigma^2)$	$E(\hat{\lambda}_c/\sigma)$	$Var(\hat{\lambda}_c/\sigma)$	$MSE(\hat{\lambda}_c/\sigma)$
$\sqrt{2}$	3	.721	.684	.761	2.963	1.556	1.557
$\sqrt{2}$	6	.389	.271	.643	6.167	.572	.600
$\sqrt{2}$	9	.355	.197	.612	9.110	.486	.498
$\sqrt{2}$	12	.476	2.575	2.850	12.042	.621	.623
$2\sqrt{2}$	3	1.144	1.067	1.088	2.771	.797	.849
$2\sqrt{2}$	6	1.277	1.857	1.934	5.794	.822	.864
$2\sqrt{2}$	9	1.254	1.286	1.350	8.901	.581	.591
$2\sqrt{2}$	12	1.333	3.513	3.624	11.884	.566	.579
$3\sqrt{2}$	3	.862	.737	.756	3.192	.733	.770
$3\sqrt{2}$	6	1.067	.746	.750	5.924	.612	.618
$3\sqrt{2}$	9	1.075	.614	.619	8.933	.463	.467
$3\sqrt{2}$	12	1.199	2.098	2.138	11.906	.576	.585
$4\sqrt{2}$	3	1.250	.758	.821	3.374	1.363	1.503
$4\sqrt{2}$	6	1.327	1.651	1.758	5.853	.862	.878
$4\sqrt{2}$	9	.923	.286	.292	8.973	.459	.460
$4\sqrt{2}$	12	1.008	.325	.325	11.949	.470	.473

12. The mean, variance, and mean square error of $\hat{\sigma}_c^2$ and $\hat{\lambda}_c$ when $k_1 = 5$.

c/σ	λ/σ	$E(\hat{\sigma}_c^2/\sigma^2)$	$Var(\hat{\sigma}_c^2/\sigma^2)$	$MSE(\hat{\sigma}_c^2/\sigma^2)$	$E(\hat{\lambda}_c/\sigma)$	$Var(\hat{\lambda}_c/\sigma)$	$MSE(\hat{\lambda}_c/\sigma)$
$\sqrt{2}$	3	1.089	1.400	1.408	2.718	1.513	1.593
$\sqrt{2}$	6	.448	.908	1.213	6.152	.578	.601
$\sqrt{2}$	9	.419	.962	1.300	9.092	.482	.490
$\sqrt{2}$	12	.521	5.406	5.635	12.043	.596	.598
$2\sqrt{2}$	3	1.195	1.632	1.670	2.774	.888	.939
$2\sqrt{2}$	6	1.343	2.353	2.471	5.836	.776	.803
$2\sqrt{2}$	9	1.375	2.768	2.909	8.903	.746	.755
$2\sqrt{2}$	12	1.459	7.303	7.513	11.902	.645	.655
$3\sqrt{2}$	3	.865	.620	.788	3.152	.614	.637
$3\sqrt{2}$	6	1.107	1.247	1.117	5.939	.636	.640
$3\sqrt{2}$	9	1.051	.414	.644	8.960	.448	.450
$3\sqrt{2}$	12	1.142	1.834	1.354	11.952	.441	.443
$4\sqrt{2}$	3	1.339	1.104	1.219	3.286	.854	.936
$4\sqrt{2}$	6	1.299	2.560	2.650	5.915	.753	.760
$4\sqrt{2}$	9	.890	.202	.212	8.991	.441	.441
$4\sqrt{2}$	12	.979	.224	.224	11.978	.412	.412

4.2 Five equation estimators

The estimators of σ^2 and λ that are proposed in this section are similar to those in the last section except that all four equations for Q_c are used. That is $\hat{\sigma}^2$ and $\hat{\lambda}$ are found so that the sum of the squared residuals,

$$R_1 = [S^2 - \hat{\sigma}^2 - k_1 k_2 \hat{\lambda}^2/n(n-1)]^2 + \sum_{j=1}^{4} [Q_{j\sqrt{2}} - g_{j\sqrt{2}}(\hat{\sigma}^2, \lambda)]^2,$$

is minimized. Once again the procedure that was used to find the values of R_1 is a combination of the Gauss-Newton method and the method of Steepest Descent for estimating the parameters of non-linear models.

Table 13 gives the mean, variance, and mean square error of $\hat{\sigma}^2$ and $\hat{\lambda}$ for various values of λ and k_1 from the Monte Carlo study.

13. The mean, variance, and mean square error of $\hat{\sigma}^2$ and $\hat{\lambda}$.

K_1	λ/σ	$E(\hat{\sigma}^2/\sigma^2)$	$Var(\hat{\sigma}^2/\sigma^2)$	$MSE(\hat{\sigma}^2/\sigma^2)$	$E(\hat{\lambda}/\sigma)$	$Var(\hat{\lambda}/\sigma)$	$MSE(\hat{\lambda}/\sigma)$
1	3	1.419	.627	.802	1.589	2.731	4.722
1	6	1.079	.485	.491	5.781	1.942	1.990
1	9	1.008	.305	.305	8.950	1.059	1.062
1	12	1.069	.352	.357	11.924	1.165	1.170
2	3	1.201	.906	.947	2.642	.961	1.089
2	6	1.100	.937	.947	5.855	1.025	1.046
2	9	1.007	.277	.277	8.972	.578	.579
2	12	1.046	.358	.360	11.967	.623	.624
3	3	.761	.490	.547	3.071	.645	.650
3	6	1.084	.704	.711	5.913	.592	.600
3	9	.962	.297	.298	8.965	.446	.447
3	12	1.045	.365	.367	11.943	.458	.461
5	3	.818	.431	.464	3.058	.509	.512
5	6	1.044	.552	.554	5.966	.441	.442
5	9	.938	.212	.216	8.984	.428	.428
5	12	1.009	.245	.245	11.974	.399	.400

Examination of Table 13 shows that $\hat{\sigma}^2$ and $\hat{\lambda}$ are approximately unbiased estimates of σ^2 and λ for all values of k_1 except when $\lambda=3$. I think that one reason that the estimators are not unbiased when $\lambda=3$ is that the values of C used were quite large with respect to this value of λ and in this case the values of Q_c were approximately the same for all c greater than $\sqrt{2}$. I believe that if the values of c were chosen smaller, that the procedure suggested here will do a good job of estimating σ^2 and λ for small values of λ as well.

5. SUMMARY AND CONCLUSIONS

The estimators V_k and V_k^* are quite simple to calculate and give a good range in which σ^2 will probably fall. However, tables for the divisors V_k have not yet been generated for any cases other than when the sample size is 10.

If a high speed computer is available, I think estimators similar to those given in Section 4.2 hold the most promise. Different methods for choosing the cs should be evaluated. Perhaps the most useful would be choosing a set of c's which are functions of the sample values such as functions of the range of the sample.

REFERENCES

1. Ascombe, F. J. (1960), "Rejection of Outliers", Technometrics 2:123-147.

2. David, H. A. and Paulson, A. S. (1965). "The Performance of Several Tests for Outliers", Biometrika 52:429-436.

3. Dixon, W. J. (1950), "Analysis of Extreme Values", Annals of Mathematical Statistics 21:488-506.

4. Dixon, W. J. (1953), "Processing Data for Outliers", Biometrics 9:74-89.

5. Grubbs, Frank E. (1950), "Sample Criteria for Testing Outlying Observations", Annuals of Mathematical Statistics 21:27-58.

6. Grubbs, Frank E. (1969), "Procedures for Detecting Outlying Observations in Samples", Technometrics 11:1-21.

7. Guttman, Irwin and Smith, Dennis E. (1971), "Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution II: Estimation of the Variance", Technometrics 13:101-111.

8. Johnson, Dallas E. and Graybill, Franklin, A. (1972), "Estimation of σ^2 in a Two-Way Classification Model with Interaction", Journal of the American Statistical Association 67:388-394.

9. McMillan, R. G. (1971), "Tests for One or Two Outliers in Normal Samples with Unknown Variance", Technometrics 13:87-100.

10. McMillan, R. G. and David, H. A. (1971), "Tests for One or Two Outliers in Normal Samples with Known Variance", Technometrics 13:75-85.

11. Tiao, G. C. and Guttman, Irwin (1967), "Analysis of Outliers with Adjusted Residuals", Technometrics 9:541-559.

12. Tietjen, Gary L. and Moore, Roger H. (1961), "Some Grubbs-Type Statistics for the Detection of Several Outliers", Technometrics 14:583-597.

13. Quesenberry, C. P. and David, H. A. (1961), "Some Tests for Outliers", Biometrika 48:379-390.